COMP9313 18S2 Project3 Optimization

z5092923
Jintao Wang

1.Prefix filter

For every record in every file, do prefix filtration first. Note this method is not exactly the same with that shown in the PPT. For a record, say that the length of the element list is ***l*** and the threshold is ***x***, we choose first ***n*** elements after sorting these elements to build inverted index where $n = l - \lceil l * x \rceil + 1$. The reason for doing this is that for this record ***R***, which have a similarity (larger than ***x***) with another record, there should be at least one element among this ***n*** elements. Compared with the algorithm in the PPT, the time complexity is higher but it is much easier to implement.

2.Length filter

When combining two files, say there is a pair (id1, id2) where id1 id from file1 and id2 is from file2, if Min (length of id1, length of id2) / Max (length of id1, length of id2) is less than threshold ***x***, even in the best case, the similarity will not be greater than or equal to ***x.***

3.Remove duplicates

After emitting all the possible pairs, it is necessary to remove redundancy before calculating the similarities.